MULTIMODAL AND CROSS-MODAL PROCESSING IN INTERACTIVE SYSTEMS BASED ON TANGIBLE ACOUSTIC INTERFACES

A. Camurri, C. Canepa, C. Drioli, A. Massari, B. Mazzarino, G. Volpe
University of Genova
DIST - InfoMus Lab
www.infomus.dist.unige.it

ABSTRACT

This paper presents some recent developments at DIST-InfoMus Lab on multimodal and cross-modal processing of multimedia data streams with a particular focus on interactive systems exploiting Tangible Acoustic Interfaces (TAIs). In our research multimodal and cross-modal algorithms are employed for enhancing the extraction and analysis of the expressive information conveyed by gesture in non-verbal interaction. The paper discusses some concrete examples of such algorithms focusing on the analysis of high-level features from expressive gestures of subjects interacting with TAIs. The features for explicit support of multimodal and cross-modal processing in the new EyesWeb 4 open platform (available at www.eyesweb.org) are also introduced. Results are exploited in a series of public events in which the developed techniques are applied and evaluated with experiments involving both experts and the general audience. Research is carried out in the framework of the EU-IST STREP Project TAI-CHI (Tangible Acoustic Interfaces for Computer-Human Interaction).

1. INTRODUCTION

The development of multimodal and cross-modal algorithms for integrated analysis of multimedia streams offers an interesting challenge and opens novel perspectives for research on multimedia content analysis, multimodal interactive systems, innovative natural and expressive interfaces [1].

Multimodal processing enables the integrated analysis of information coming from different multimedia streams (audio, video) and affecting different sensorial modalities (auditory, visual). Cross-modal processing enables exploiting potential similarities in the approach for analyzing different multimedia streams so that algorithms developed for analysis in a given modality (e.g., audio) can be also employed for analysis in another modality (e.g., video).

In the EU-IST Project TAI-CHI (Tangible Acoustic Interfaces for Computer-Human Interaction), we face multimodal and cross-modal processing. The aim is to enhance extraction and analysis of information related to the high-level expressive content in communication. The focus is on non-verbal gestural interaction using systems based on Tangible Acoustic Interfaces (TAIs).

Designing and developing TAIs consists of exploring how physical objects, augmented surfaces, and spaces can be transformed into tangible-acoustic embodiments of natural seamless unrestricted interfaces. The ultimate goal of the TAI-CHI project is therefore to design TAIs employing physical objects and space as media to bridge the gap between the virtual and physical worlds and to make information accessible through large size touchable objects as well as through ambient media. In this framework, a relevant aspect for the success of TAI-based interactive systems is their ability of processing expressive information. In order to provide systems with such skill, research needs to address the development of models and algorithms for multimodal and cross-modal high-level analysis and interpretation of integrated data extracted from video images, acoustic tangible interfaces, and acoustic localization systems. Interpretation of gesture includes expressive content.

This paper presents and discusses some concrete examples of multimodal and cross-modal algorithms we used for analysis of expressive gesture [2] during interaction with TAIs. The algorithms have been implemented as software modules (blocks) or applications (patches) for the new EyesWeb 4 platform [3] (www.eyesweb.org) that differently from its predecessors directly and explicitly supports multimodal processing. The main features of the new EyesWeb 4 platform related to the support of multimodality and cross-modality are also briefly introduced.

2. CROSS-MODAL PROCESSING: VISUAL ANALYSIS OF ACOUSTIC PATTERNS

A first application of cross-modal processing consists of the analysis by means of computer vision techniques of acoustic patterns extracted from an audio signal by means of a collection of EyesWeb 4 modules for auditory modeling.

Such modules are included in an EyesWeb library providing the whole auditory processing chain, i.e., cochlear filter banks, hair cell models, and auditory representations including excitation pattern, cochleogram, and correlogram [4]. The design of the cochlear filter banks relies on the Matlab Auditory Toolbox [5]. To date, a filter bank

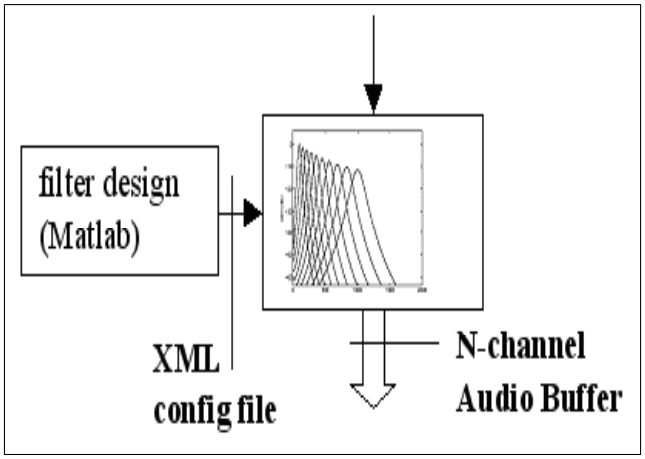

Figure 1. Design of the auditory filter bank through the Matlab Auditory Toolbox

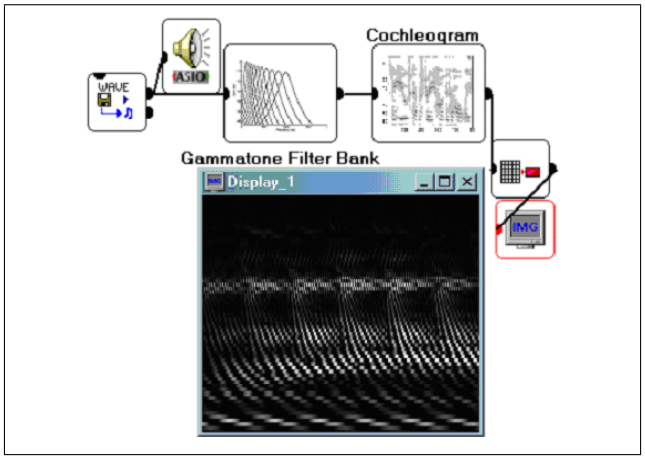

Figure 2. Cochleogram of a voice sound obtained through the auditory model blocks

configuration can be exported in XML format and loaded into the EyesWeb plug in (see Figure 1). For example the cochleogram of voice sound is depicted in Figure 2.

The cochleogram images can be analyzed by image processing techniques to extract information that is not so directly accessible through audio analysis (e.g., activation of particular regions in the image, pattern matching with template images).

In a first application of cross-modal techniques we analyzed the cochleogram images by applying to them the techniques for motion analysis included in the EyesWeb Gesture Processing Library [6]. For example, in order to quantify the variation of the cochleogram, i.e., the variance over time of the spectral components in the audio signal, we used Silhouette Motion Images (SMIs) and Quantity of Motion (QoM) [1]. Figure 3 show the SMI of a cochleogram (red shadow). It represents the combined variation of the audio signal over time and frequency in the last 200 ms. The area (i.e., number of pixels) of the SMI (that in motion analysis is usually referred to as Quantity of Motion, i.e., the amount of detected overall motion) summarizes such variation of the audio signal, i.e., it can

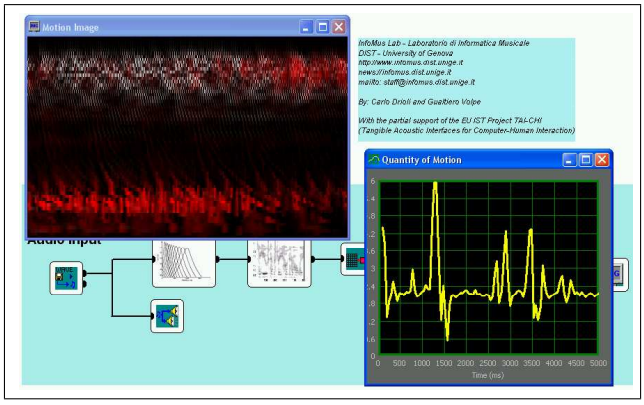

Figure 3. SMI of a cochleogram (red shadow) and graph of the corresponding QoM

be considered as the detected amount of variation of the audio signal both along time and along frequency in the time interval over which the corresponding SMI is computed (200 ms in this example).

From a first analysis of the data obtained with this approach it seems that the QoM obtained from the SMIs of the cochleograms can be employed for onset detection especially at the phrase level, i.e., it can be used for detection of phrase boundaries. In speech analysis the same technique can be used for segmenting words. Current research includes performance analysis and comparison with state-of-the-art standard techniques.

3. CROSS-MODAL PROCESSING: AUDITORY-BASED ALGORITHMS FOR MOTION ANALYSIS

Cross-modal processing applications can also be designed in which the analysis of movement and gestures is inspired by audio analysis algorithms. An example is the patch shown in Figure 4, in which a pitch detector is used to measure the frequency of periodic patterns in human gestures: the vertical displacement of a moving hand, measured from the video input signal and rescaled, is converted into the audio domain through an interpolation block, and then analyzed through a pitch detector based on the autocorrelation function.

Motion-derived signals and audio signals differ in terms of sampling rate and band characteristics. The conversion from a motion-derived signal to one in the audio domain can be performed in principle by upsampling and interpolating the input signal, and a dedicated conversion block is available to perform this operation. If m_{i-1} and m_i are the previous and present input values respectively, and t_i is the initial time of the audio frame in seconds, the audio-rate samples are computed by linear interpolation as

$$s(t_i + \frac{n}{F_s}) = m_{i-1} + n\frac{(m_i - m_{i-1})}{N_s}, n = 1 \ldots N_s$$

where N_s is a selected audio frame length at a given audio sampling rate F_s. However, often sound analysis algorithms are designed to operate in frequency ranges that

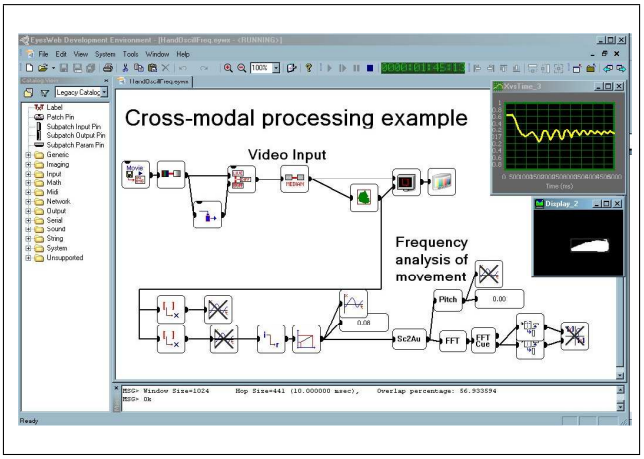

Figure 4. An example of EyesWeb application for cross-modal analysis of movement: the hand vertical displacement, measured from the video signal, is converted into the audio domain and analyzed through a pitch detector.

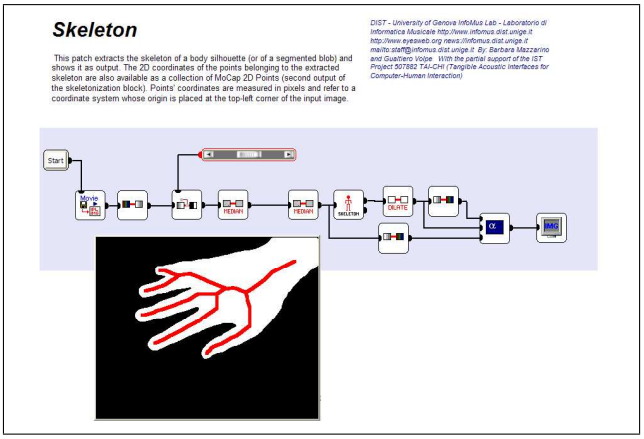

Figure 5. An EyesWeb 4 patch extracting the skeleton of a hand touching a TAI

are much higher if compared to those related to the velocity of body movements. For this reason, the conversion block also provides amplitude modulation (AM) and frequency modulation (FM) functions to shift the original signal band along the frequency axis. If

$$c(t) = A_c \cos(2\pi f_c t)$$

is a sinusoidal carrier wave with carrier amplitude A_c and carrier frequency f_c, an AM audio-rate signal can be computed as

$$s_m(t) = A_c s(t) \cos(2\pi f_c t),$$

and an FM signal as

$$s_m(t) = A_c \cos(2\pi f_c t + 2\pi \int_0^t s(t)dt).$$

The approach to motion analysis by algorithms inspired to acoustic and/or musical cues extraction can be explored further. A possible application is, for example, the control of a digital score reproduction (e.g., a MIDI file) through the detection of tempo, onset, IOI, and other similar musical parameters from the arm and hand movements.

4. MULTIMODAL PROCESSING FOR ANALYSIS OF TOUCH GESTURES

As an example of multimodal analysis of gestural information we consider a first experimental TAI application for the analysis of touch gesture. The aim is twofold: (i) locate where on the TAI the touch gesture takes place, and (ii) analyze how touching is performed (i.e., individuating the expressive qualities of the touching action, such as for example whether the touching action is light and delicate or heavy and impulsive).

The approach to analysis is multimodal since we use both information extracted from the acoustic signal generated by the touching action on the TAI and from the information extracted from a video-camera toward the touching position.

Localization is based on two algorithms for in-solid localization of touching positions developed by the partners in the TAI-CHI project. The first algorithm, developed by the Image and Sound Processing Group at Politecnico di Milano employs 4 sensors and is based on the computation of the Time Delay of Arrival (TDOA) of the acoustical waves to the sensors [7]. The second algorithm developed by the Laboratoire Ondes et Acoustique at the Institut pour le Developement de la Science, l'Education et la Technologie, Paris, France, employs just 1 sensor and is based on pattern matching of the sound patterns generated by the touching action against a collection of stored patterns. In order to increase the reliability of the detected touching position we developed an EyesWeb application integrating the two methods and compensating the possible weakness of one method with the outcomes of the other one.

The position and time of contact information obtained from audio analysis can be employed to trigger and control in a more precise way the video-based gesture analysis process: e.g., we are testing hi-speed and hi-res videocameras in EyesWeb 4 in which it is also possible to select the portion of the active ccd area using (x,y) information from a TAI interface.

Video-based analysis (possibly combined with information extracted from the sound generated by the touching action, e.g., the sound level) is then used for extraction of expressive qualities. Gesture analysis is based on hand detection and tracking and builds upon the extraction of information concerning both static and dynamic aspects. As for the static aspects we developed a collection of EyesWeb modules for real-time classification of hand postures. Classification employs machine learning techniques (namely, Support Vector Machines). As for the dynamic aspects we used the expressive features currently available in the EyesWeb Expressive Gesture Processing Library (e.g., Quantity of Motion, Contraction/Expansion, Directness Index etc.). Figure 5 shows for example the output of an EyesWeb module for the extraction of the hand skeleton.

Figure 6. A chair equipped with sensors is transformed in a Tangible Acoustic Interface (TAI). This chair will be used in the music theater work"Un avatar del diavolo" (composer Roberto Doati), La Biennale di Venezia, September 30th, 2005.

In other words, while the contact position is detected through an acoustic based localization system, visual information is employed to get information on how the hand approaches and touches the interface (e.g., with a fluent movement, or in a hesitating way, or in a direct and quick way etc.).

5. APPLICATIONS

An EyesWeb application for multimodal analysis of touch gesture was employed in the music theater work "Un avatar del diavolo" at La Biennale in Venice (composer Roberto Doati) on September 30, 2005. This application was used for detecting the touching position on everyday life objects such as chairs and tables. In particular, the touching gestures performed by an actor on a chair (see Figure 6) were analyzed and classified. The extracted features and the recognized gesture class were used for real-time generation and processing of visual and audio content according to the composer's needs.

Interesting HCI issues emerged from this artistic project: interaction design issues, taxonomies of interaction gestures needed to fully exploit a chair as a TAI.

Another experimental TAI application we are developing consists in a method for performing acoustic handwriting (see Figure 7), i.e., for recognizing a drawn symbol from the sound produced by a pen on a TAI. This application integrates analysis of acoustic signals with pattern matching and machine learning techniques.

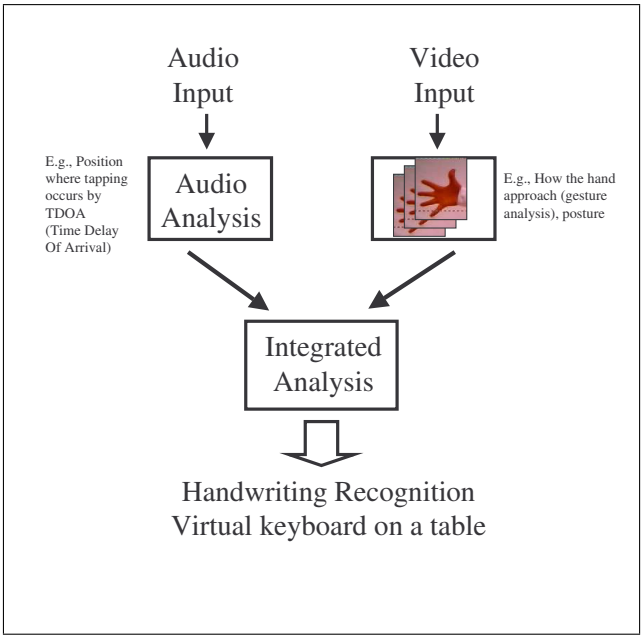

Figure 7. Multimodal analysis of hand gesture for TAI applications such as acoustic handwriting and virtual keyboard.

6. THE EYESWEB 4 PLATFORM FOR MULTIMODAL AND CROSS-MODAL PROCESSING

The need for fully integrated and supported multimodal processing of data streams from several channels (e.g., visual, auditory) led us to a complete redesign of the EyesWeb open platform: the new EyesWeb version 4 [3]. A relevant aspect in the new EyesWeb version 4 platform is the explicit support to multimodality and cross-modality in the EyesWeb language. This is obtained through a new kernel, now providing (i) low-level scheduling mechanisms for managing different data streams (e.g., auditory and visual data) at different sampling rates, and (ii) high-level extensions toward integration of gesture and audio processing, aimed at real-time analysis of expressive information.

Concerning the low level support, besides the basic requirement to manage several datatypes (e.g., audiovisual and sensor systems streams) in a common environment, new features have been added. One of such features supports the automated transformation of datatypes of different, but compatible, domains. This is particularly useful to verify the effectiveness of an algorithm originally designed and implemented for a very specific domain. As an example, an FFT block [1] working on matrices can be easily used to work on audio streams, since the conversion of audio buffers to matrices is automatically added by the system, without the need for any explicit action

[1] A block is a software module for the EyesWeb platform

on the user side. Another important feature is the possibility to design and develop blocks working on a whole family of datatypes. Where the previous version of EyesWeb could distinguish among specific blocks (working on a given datatype) or general purpose blocks (working on all datatypes), this new version supports the generalization of the block to the datatypes sharing a set of specified characteristics. This enables the development of blocks working on homogenous sets of datatypes, without the need to know them in advance, thus, it does not limit the expansibility of EyesWeb. Referring to the above FFT example, a better designed block could exploit such feature and work natively on both the audio buffer and the matrix datatypes, as they share core common characteristics (i.e., they both implement a common interface). This approach has the further advantage by a performance point of view, since it avoids consuming processing power for datatype conversion.

Cross-modal and multimodal processing is supported by several other features, including the timestamping of datatypes, which has been greatly enriched in this new version of EyesWeb. At a low-level, each datatype is associated with a set of timestamps that the kernel can use to synchronize the data according to different needs. At a higher-level the EyesWeb kernel can be flexibly extended by including novel specific and customized scheduling and synchronization mechanisms managing flow of datatypes and activation of blocks.

7. CONCLUSIONS AND FUTURE WORK

This paper presented a collection of concrete examples of cross-modal and multimodal analysis techniques for non-verbal expressive gesture processing applied to data streams from Tangible Acoustic Interfaces. The preliminary results from these sample applications indicate the potentialities of a multimodal and cross-modal approach to expressive gesture processing: cross-modal techniques enable to adapt to the analysis in a given modality approaches originally conceived for another modality, allowing in this way the development of novel and original techniques. Multimodality allows integration of features and use of complementary information, e.g., use of information in a given modality for supplementing lack of information in another modality or for reinforcing the results obtained by analysis in another modality.

While these preliminary results are encouraging, further research is needed for fully exploiting cross-modality and multimodality in expressive gesture processing. For example, an open problem which is currently under investigation at our Lab concerns the development of high-level models allowing the definition of cross-modal features. That is, while the work described in this paper concerns cross-modal algorithms, a research challenge consists of identifying a collection of features that, being at a higher-level of abstraction with respect to modal features, are in fact independent of modalities and can be considered cross-modal since they can be extracted from and applied to data coming from different modalities. Such cross-modal features are abstracted from the currently available modal features and define higher-level feature spaces allowing for multimodal mapping of data from one modality to another.

8. ACKNOWLEDGMENTS

We thank our colleagues at DIST-InfoMus Lab. This research is partially supported by the EU-IST Project TAI-CHI (Tangible Acoustic Interfaces for Computer-Human Interaction).

9. REFERENCES

[1] Camurri, A., Mazzarino, B., Volpe, G. "Expressive interfaces", *Cognition, Technology and Work*, 6(1):15-22, Springer-Verlag, 2004.

[2] Camurri A., Mazzarino B., Ricchetti M., Timmers R., Volpe G. "Multimodal analysis of expressive gesture in music and dance performances", in A. Camurri, G. Volpe (Eds.), *Gesture-based Communication in Human-Computer Interaction*, LNAI 2915, pp. 20-39, Springer Verlag, 2004.

[3] Camurri A., De Poli G., Leman M., Volpe G. "Toward Communicating Expressiveness and Affect in Multimodal Interactive Systems for Performing Art and Cultural Applications", *IEEE Multimedia Magazine*, 12(1):43-53, IEEE CS Press, 2005.

[4] Camurri A., Coletta P., Drioli C., Massari A., Volpe G. "Audio Processing in a Multimodal Framework", *Proceedings 118th AES Convention*, Barcelona, Spain, 2005.

[5] Slaney M. "Auditory toolbox documentation. Technical report 45". Technical report, Apple Computers Inc., 1994. Available for download at: http://www.slaney.org/malcolm/pubs.html

[6] Camurri, A., Mazzarino, B., Volpe, G. "Analysis of Expressive Gesture: The EyesWeb Expressive Gesture Processing Library", in A. Camurri, G. Volpe (Eds.), *Gesture-based Communication in Human-Computer Interaction*, LNAI 2915, pp. 20-39, Springer Verlag, 2004.

[7] Polotti P., Sampietro M., Sarti A., Tubaro S., Crevoisier A. "Acoustic Localization of Tactile Interactions for the Development of Novel Tangible Interfaces", *Proceedings of the 8th Int. Conference on Digital Audio Effects (DAFX-05)*, Madrid, Spain, 2005.